COTS Parallel Archive Integration Experiences

Gary Grider, HPC-DO; Hsing-Bung Chen, HPC-5; Cody Scott, Milton Turley,
Aaron Torres, Kathryn Sanchez, John Bremer, HPC-3

There is demonstrated need for current and future archive storage systems for high performance computing to 1) scale to very high bandwidths, 2) scale in metadata performance, 3) support policy-based hierarchical storage management capability, 4) scale in supporting changing needs of very large data sets, 5) support a standard interface, and 6) use commercial-off-the-shelf (COTS) hardware. This is akin to changes in parallel file systems, although file systems perform at one or more orders of magnitude faster. Archive systems continue to move closer to file systems in their design due to the need for speed and bandwidth, especially metadata searching speeds, by using more caching and less robust semantics. Currently the number of extreme highly scalable parallel archive solutions is very small, especially those that will move a single large highly striped parallel disk file to many tapes in parallel. We are developing a hybrid storage approach, using COTS components and innovative software technology to bring new capabilities to productive use for the HPC community much faster than if we create and maintain a complete end-to-end unique parallel archive software solution.

In this project we worked to integrate a COTS global parallel file system and a standard backup/archive product with a very small amount of additional user space code to provide a scalable and parallel solution that overlaps highly in function with current niche parallel archive product(s), including 1) doing parallel movement to/from tape for a single large parallel file, 2) hierarchical storage management, 3) Information Lifecycle Management (ILM) features, 4) high volume (nonsingle parallel file) archives for backup/archive/content management, and 5) leveraging all free file movement/management tools in Linux such as copy, move, ls, or tar.

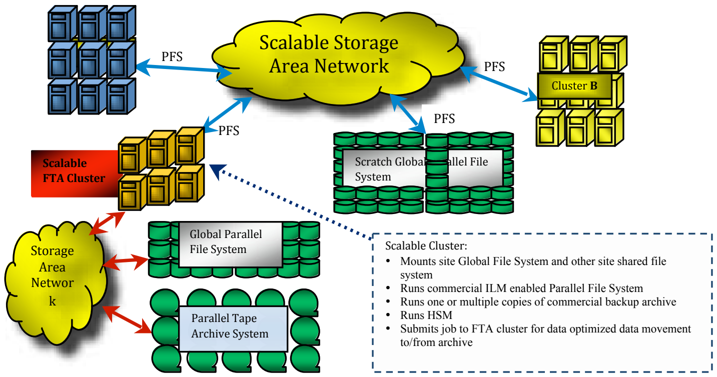
The Advanced Simulation and Computing (ASC) program at LANL has a goal to pilot a more COTS-based archive in less than 5 years. We had an opportunity to do a pilot archive project for the LANL Roadrunner cluster while it was in its testing phase. We chose the IBM General Parallel File system (GPFS) for the parallel file system because of the new ILM features. We chose the IBM Tivoli Storage Manager (TSM) because we had it in-house. We have designed, developed, and integrated the following features in the proposed COTS Parallel Archive System (Fig. 1, Fig.2):

- A parallel tree walker and copy user space utility
- A storage pool (stgpool) support in utility (using file system application program interface (API))
- An efficient ordered retrieval in utility (using dmapi API and back-end tape system query)
- Archive Parallel File System to support ILM stgpool features
- Archive back-end to support ILM stgpool and colocation features
- A Filesystem in UserSpace (FUSE) to break up enormous files into parts that can be migrated and recalled in parallel to/from back-end tape system

We show our proposed COTS Parallel Archive System in two parts: the back-end system (Fig. 3) and the front-end system (Fig. 4a and b).

We have built and applied our working COTS Parallel Archive System to one of the world's fastest computer systems, Roadrunner. We have successfully supported the seven Roadrunner Open Science Projects to archive over 1 petabyte of data and demonstrated its capability to address requirements of future archive storage systems. This COTS system is now providing archive services in LANL's Turquoise open collaboration network.

For more information contact Cody Scott at cscott@lanl.gov.


Scalable Storage Area Network
PFS
PFS
PFS
PFS
Cluster B
Scalable FTA Cluster
Scratch Global Parallel File System
Storage Area Network
Global Parallel File System
Parallel Tape Archive System

Scalable Cluster:

- Mounts site Global File System and other site shared file system
- Runs commercial ILM enabled Parallel File System
- Runs one or multiple copies of commercial backup archive
- Runs HSM
- Submits job to FTA cluster for data optimized data movement to/from archive

Fig. 1. Proposed COTS parallel archive system to deploy a parallel archive with a parallel file system as its first tier of storage.

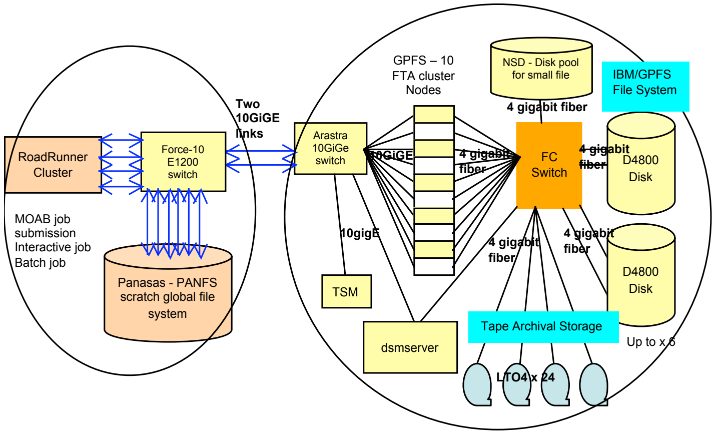


Fig. 2. COTS parallel archive system for LANL's Open Science Projects.

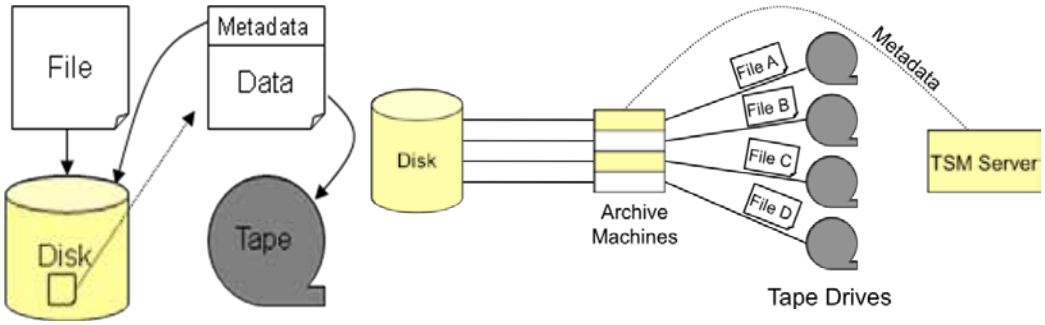
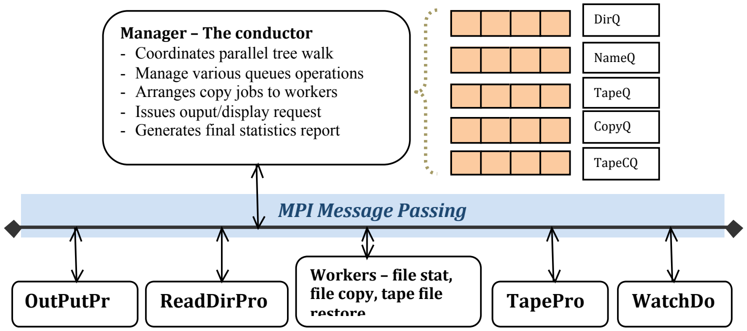


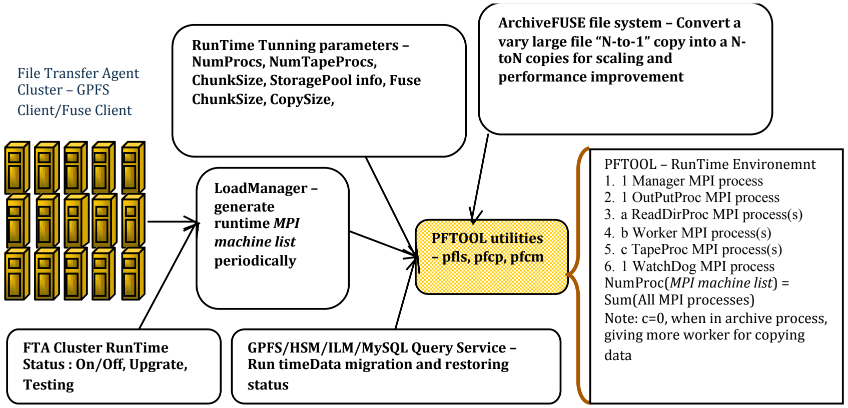
Fig. 3. The backend system of the COTS parallel archive system.



Manager – The conductor
- Coordinates parallel tree walk
- Manage various queues operations
- Arranges copy jobs to workers
- Issues ouput/display request
- Generates final statistics report

Workers – file stat, file copy, tape file restore

Fig. 4a. PFTool software system diagram.


File Transfer Agent Cluster – GPFS Client/Fuse Client
PFTOOL utilities – pfls, pfcp, pfcm

RunTime Tunning parameters – NumProcs, NumTapeProcs, ChunkSize, StoragePool info, Fuse ChunkSize, CopySize,

ArchiveFUSE file system – Convert a vary large file "N-to-1" copy into a N-toN copies for scaling and performance improvement

LoadManager – generate runtime *MPI machine list* periodically

FTA Cluster RunTime Status : On/Off, Upgrate, Testing

GPFS/HSM/ILM/MySQL Query Service – Run timeData migration and restoring status

PFTOOL – RunTime Environemnt
1. 1 Manager MPI process
2. 1 OutPutProc MPI process
3. a ReadDirProc MPI process(s)
4. b Worker MPI process(s)
5. c TapeProc MPI process(s)
6. 1 WatchDog MPI process

NumProc(*MPI machine list*) = Sum(All MPI processes)

Note: c=0, when in archive process, giving more worker for copying data

Fig. 4b. PFTOOL system runtime operation diagram.

Funding Acknowledgments
LANL Institutional Computing Program